# Testing the Transferability of the Defense-GAN Technique against Adversarial Examples that have Undergone Physical Transformation

Rohan N. Pradhan Bibin K. Sebastian Mihir Pathare Chamanpreet Kaur Karthik Raja

***Abstract*—This paper will discuss the transferability of state of the art defense techniques for adversarial examples for deep learning systems in the physical domain. The paper explores using adversarial attacks using the Fast Gradient Sign Method (FGSM), Carlini & Wagner (CW) and DeepFool attacks to generate adversarial images that are given to the classifier as a digital and physically transformed image. Furthermore, we present novel results demonstrating the effectiveness of the state-of-the-art Defense-GAN technique to create reconstructions of images, that have undergone the physical transformation, with a significant portion of the adversarial noise filtered out. We also show, that for finer adversarial attacks, that the physical transformation itself causes a high degree of adversarial destruction, bringing to question the need for additional defenses.**

***Index Terms*—Adversarial Examples, Fast Gradient Sign Method, DeepFool, Carlini & Wagner, Generative Adversarial Networks**

## I. Introduction

Within recent years the thriving success of machine learning has led to transforming innumerable aspects of our lives. Increased availability of enormous data sets and advances in machine learning algorithm from naive approaches to broad spectrum applicability, has resulted in the fostering of machine learning's success. But along this rise came the infallibility of machine learning models to adversarial attack. Adversarial inputs are precisely perturbed to cause the model to make a mistake. Additionally, adversarial examples crafted for one model are often misclassified by other models too. This puts a big question mark on the actual learning achieved by these cutting edge machine learning models.

The significant potential of machine learning has made it possible for a viable shift from research labs to production use in public and private sectors. Deep neural networks (DNNs) are already being deployed in systems like autonomous vehicles, robotic surgery, detection of credit card fraud, maritime surveillance, air traffic control and many more mission critical fields. Therefore, the autonomous decisions made by these systems, within mission critical environments, will have a profound effect on the lives of the people interacting with it. Additionally, many mission critical systems reside in the cyber-physical domain. Since these systems are operating in the real world, where surroundings are not static and tranquil, the robustness of these cyber-physical systems is debatable, under the circumstances when adversaries interacts with these machine learning system and erroneous outputs will lead to potentially catastrophic outcomes. Despite the threat probability and damage to human lives presented by these vulnerabilities to machine learning systems being used in the cyber-physical domain, the majority of recent and past developments in arena of defense techniques give more priority to the digital domain rather than the physical world.

All these contentions in the cyber-physical domains raise an emerging concern regarding whether state of the art defense techniques for digital platforms generalize well to the physical domain. As demonstrated in Kurakin et al. [1] paper, adversarial examples are misclassified even by simply printing these examples and feeding the images taken using a ordinary camera phone to the image classification system. This instigates precarious scenarios when the adversary might be successful in attacking a model operating in real world with negligible access to its underlying structure like autonomous vehicles that have deep learning systems that use cameras/vision systems to perceive the world.

Specifically, we hope our work can positively affect the current body of working surrounding self-driving cars/autonomous vehicles. Autonomous vehicles have the potential to disrupt our preconceived notions of traffic flow stability and throughput. As described in Talebpour et al. [2], the real time sensor data provided from connected autonomous vehicles allows for real time traffic flow, patterns, and stability at the edge, which was not previously possible. Additionally, with humans not having to actively operate an autonomous vehicle, we believe that there will large increase in advertising and entertainment on roads. Most autonomous vehicles rely on vision based systems to perceive the world, including object detection and traffic signage classification. With the increase in digital traffic signs (to allow for more dynamic traffic flows) and vehicles with digital signs for advertising, there is an increased threat space for malicious actors to cause harm. For a malicious actor to maximize the adversarial impact, in this setting, we expect him/her to add adversarial noise to digital traffic signs or digital displays on vehicles. These adversarial displays will be perceived by the computer vision system of the vehicle (thereby undergoing a physical transformation). Therefore, we plan on exploring the use of digital displays/monitors to perform the physical transformation on adversarial examples to best replicate this scenario.

Our work illustrates the refurbishing of adversarial examples in physical world. We utilize attacks like FGSM, DeepFool

and CW to generate adversarial examples for a classifier and orchestrate an in depth analysis of classifier's accuracy when the same adversarial example undergo a physical transmutation. Further we examine defense techniques applicable to withstand adversarial behaviour. Therefore, we select Samangouei et al. [3] paper as a baseline, where we re-implement capabilities of GANs to defend against adversarially perturbed input images. Finally after these accomplished outcomes, we test defense mechanism for GANs against physical adversarial examples and provide novel results to make conclusions.

## II. Related Work

### A. Physical Adversarial Examples

Recently, there has been an increase in work surrounding adversarial physical examples. Specifically, the goal of the Kurakin et al. [1] paper was to investigate whether adversarial images are still able to successfully trick deep learning models after going through a physical image transformation. The paper discussed three different methods of generating adversarial images - 1). Fast Gradient Sign Method (FGSM), 2). Basic Iterative Method, and 3). Iterative Least Likely Method. The experiment described in the paper, involves printing out adversarial images and then taking photographs of the image using a mobile phone. The image is then automatically cropped to a set size, before being fed into the deep learning model. Because of the other external parameters and noise from the physical transformation (degradation of image quality through printing, non-perfect camera focus, and movement of the camera from human error), there is a significant decrease in the adversary's ability to negatively impact the classifier; the paper describes this with the term adversarial destruction. The paper showed that through physical transformation, the Fast Method is the most resilient to adversarial destruction, while more finer/granular attacks do not penetrate as well after undergoing physical transformation.

More concerningly, there has been a recent body of work addressing adversarial examples in the physical world, specifically, trying to actively reduce the adversarial destruction, thereby making the adversarial noise transcend the physical transformation. In particular, Chen et al. [4] explores modifying CNN object detectors to be more robust against physical adversarial examples. The paper utilizes the Expectation over Transformation technique to show that it is possible to generate physical adversarial transformation on stop signs that are consistently misclassified by R-CNN object detectors. This body of work further emphasizes the need for practical and efficient defense mechanisms against adversarial attack on deep learning algorithms, specifically in the physical domain.

### B. Defense Against Adversarial Examples

Meng & Chen [5] introduced a new way of defending against adversarial attacks by using auto-encoders as a filter before the image classifier. When an ensemble of auto-encoders were used, one of them was chosen randomly and this method helped in deterring the adversary from successful attacks. Auto-encoders are trained on normal examples and acts as a bottleneck which gets rid of the adversarial perturbations on the forward pass itself. On top of this novel idea, the authors proposed a detector network that will compare the test image to the normal image distribution and if the similarity is less than a threshold, it rejects the image right away. However, these methods were not full proof and as per Lu et al. [6], MagNet fails miserably with Elastic-net Attack on DNNs (EAD) attack. The EAD attack uses a hybrid of L1 and L2 distortion metrics.

On a similar fashion, Samangouei et al. [3] described a new defense mechanism against adversarial attacks on classifiers through their paper titled 'Defense-GAN: Protecting classifiers against adversarial attacks using generative models'. They proposed utilizing Generative Adversarial Networks (GAN), which are usually used for training generative models for an unknown distribution, but have a natural adversarial interpretation. The paper proposed to be an improvement over the use of auto-encoders as choke point against adversarial inputs. The authors suggested to replace the auto-encoders with the generator part of a trained Generative Adversarial Network. The adversarial input gets projected to the range of Generator and gets optimized to resemble the input image, closely, using a gradient descent based optimization algorithm that scopes the input sample space using a number of random restarts. The effect of this projection technique is to filter out adversarial perturbations that the Generator was not trained on. In particular, the method is shown to be robust against FGSM, RAND+FGSM, and CW white-box attacks. The comparisons provided in the paper portrayed results that surpasses similar defenses such as MagNet.

We undertook thorough literature review and found that there have been enormous research interest in defending against adversarial inputs. However, not many research papers described the effect of physical transform on these defense mechanisms or provide a clear comparison against the performance of these defense mechanisms against the adversarial destruction caused during physical world to digital transformation. This was our motivation to pursue the transferability of Defense-GAN to physical adversarial examples.

## III. Experimental Setup

### A. Neural Network Classifier and Dataset

We run 3 broad experiments in our work. First, we generate physical adversarial examples (using different attacks), and compare the classifier accuracy to their digital adversarial counter parts. This replicates the Kurakin et al. [1] paper. Second, we create GAN reconstructions of digital adversarial examples, that attempt to filter out the adversarial noise, and submit these reconstructions to the classifier and compare the classifier accuracy to the non-reconstructed/filtered images. This replicates the Samengoui et al. [3] paper. Finally, we combined both these papers, by creating GAN-reconstructions of the physical adversarial examples and submitting these reconstructions to the classifier. Due to this setup, we held the classifier and dataset used in all these experiments constant.

TABLE I
CONVOLUTIONAL NEURAL NETWORK CLASSIFIER ARCHITECTURE

| Layer | Description |
|---|---|
| Input Layer | Conv- 64 filters of size (5,5), Input image size(64x64x3) |
| Hidden 1 | Conv- 64 filters of size (5,5), ReLU Activation |
| Hidden 1a | Dropout 25% |
| Hidden 2 | Fully Connected with 128 Neurons, ReLU Activation |
| Hidden 2a | Dropoout 50% |
| Final Layer | Fully Connected with 2 Neurons, Softmax Activation |

*1) CelebA Dataset:* The dataset that was used for this experiment is the CelebFaces Attributes dataset (CelebA) (Liu et al. [7]). The CelebA dataset is a large-scale face dataset consisting of more than 200,000 face images having image resolution of 178x218 pixels. The RGB images were center-cropped and resized to 64x64 while used to train the GAN. The data set was split to 90% train and 10% test sets and it was used to train the GAN as well as the classifier.

*2) Virtual Machine Setup:* The Experiments were computationally intensive, therefore, we utilized a number of Google Cloud Virtual Machines to train the GAN, Classifier and run the reconstructions. The VM Configuration we used had 2xCPU with 13GB RAM and 1x Tesla P4 GPU (8GB DDR5 RAM).

### *B. Physical Adversarial Examples*

In our work, one of the primary goals is to reproduce Kurakin et al. [1] paper and then use it to extend the Samangouei et al. [3] paper. As mentioned in the Kurakin et al. [3] paper, we initially recreated the Fast Gradient Sign Method (FGSM) attack as a baseline for maximum adversarial destruction. Instead of using the iterative methods described in the Kurakin et al. [1] paper, we felt the urge to test the adversarial effects under physical transformation with more complex, sophisticated algorithms like Carlini and Wagner (CW) and DeepFool attacks.

After creating the adversarial examples, using the FGSM, CW, and DeepFool attacks, the physical transformations of these adversarial examples were generated. Unlike the Kurakin et al. [1] paper, we did not use printed images as the physical transformation. As described earlier, we aim for our work to be relevant in the safety and security of the deep learning systems used in autonomous vehicles. Therefore, instead of printing out adversarial images we used a webcam to capture the adversarial images that were being displayed on a LED monitor. We chose this to closely replicate the deep learning based vision based systems used in autonomous cars to perceive the world, including digital signage or adversarial examples on digital displays.

Initially, when generating physical adversarial examples, we had selected a low specification Microsoft webcam. This resulted in extremely poor results. The resulting images from the Microsoft webcam were very out of focus and extremely over exposed. We discovered that this webcam did not have the necessary granular controls to adjust the ISO, auto-focus, and anti-flicker speed. Instead, we opted for a better webcam from Logitech. The materials and parameters used are described below:

TABLE II
CAMERA SETTINGS

| Info | Description |
|---|---|
| Camera Used | Logitech C920 |
| Flicker Reduction | 60Hz |
| Distance to Monitor | 5cm |
| Autofocus | Off |

TABLE III
MONITOR SETTINGS

| Info | Description |
|---|---|
| Monitor Used | Dell SE2416hx 23.8" LED |
| Monitor Brightness | 40% |
| Monitor Contrast | 60% |
| Monitor Resolution | 1920x1080px |

Our results are described below. We first show the model classifier accuracy against digital adversarial examples using FGSM, CW, and DeepFool attacks. We then show the model classifier accuracy against the physical adversarial examples using the same attacks.

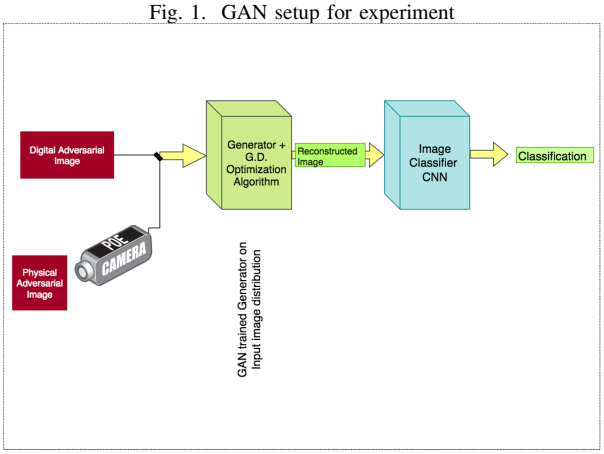
Utilizing our experimental setup described above, as a baseline we first evaluated the classier on set of clean images and attained an accuracy of 100%. Then we drafted digital adversarial images by deploying FGSM, CW and DeepFool attacks. Thereby passing these adversarial crafted images to our image classifier system, we congregated the results of classifier accuracy under separate attack scenarios. $\epsilon$ value of 0.07 for FGSM attack helped us achieve zero percent classifier accuracy. Similarly we succeeded in achieving zero percent accuracy for CW and deep fool attacks. Moving forward to generate the physical transformation of these digital adversarial images, we deployed the webcam to capture the adversarial images corresponding to distinct attacks respectively. Then we compiled the results of classifier accuracy for physical transformation settings. $\epsilon$ value of 0.07 for FGSM attack gave classifier accuracy of 10% after physical transformation. And for CW and DeepFool attacks we achieved an accuracy of 80% and 87% respectively. Refer to Table IV and V for a summary of results. Please refer to the Appendix for a detailed table.

### *C. Defense-GAN*

Generative Adversarial Networks (GAN) are an interesting idea that were introduced by a group of researchers lead by Ian J. Goodfellow [8]. The main idea behind a GAN is to have two competing neural network models. One takes random numbers as input and generates images (and so is called the generator $G$). The other model (called the discriminator $D$) receives images from both the generator and the training data, and should be able to distinguish between the two sources. These two networks play a continuous game, where the generator is learning to produce more and more realistic

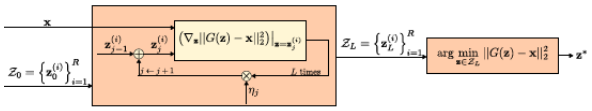
samples, and the discriminator is learning to get better and better at distinguishing generated data from real data. These two networks are trained simultaneously, and the goal is to have the generated samples to be indistinguishable from real data by continuous improvement. The loss function which was used in the initial GAN, introduced by Ian Goodfellow, suffered from issues such as vanishing gradients, failure to converge and mode collapse. Researchers have done extensive work to improve trainability of GANs and a variant called Wasserstein GAN [9] is used in our project. This paper utilizes Earth Mover distance (Wasserstein distance) as the metric to compare real and GAN generated image distributions.

Fig. 1. GAN setup for experiment




Samangouei et al. [3] proposed to use a train GAN on real images and use the Generator during the image classification step as a filter before the image classifier. Unlike what is proposed in Meng & Chen [5] paper, where they use an auto-encoder to reconstruct the input image, the Generator in the GAN does not accept images in its input layer. Samangouei et al. [3] came up with a novel optimization algorithm that creates a reconstruction of the the input image which is very close to it. The algorithm has steps as follows. Given an input image $x$ (which has been adversarially perturbed), project $x$ onto the range of $G$ by solving the minimization problem $z^* = argmin||G(z) - x||^2$ to find the optimum $z^*$. This optimization step iterates through $L$ steps of Gradient Descent and $R$ random restarts in the $z$ input space. Each Gradient Descent step tweaks the $z$ vector towards a reconstructed G(z) which is closer to $x$. $R$ random restarts ensure that Generator gets to scope the input latent space and get out of local minima if any. The resulting $x^* = G(z^*)$, which is hypothesized to be a very close representation of the input image without the adversarial noise, gets input to the classifier trained on the true distribution. In particular, the defense can be applied in conjunction with any classifier (including already adversarial trained classifiers), and does not assume any specific attack model.

Fig. 2. GAN optimization algorithm as described in Defense-GAN



The authors of the paper generously uploaded their work on Github. The code was comprehensive with a pipeline that downloads the dataset, trains the GAN, trains the image classifiers and tests the classifiers with different modes of defense mechanisms in place. On the paper, the illustrations were mostly done on MNIST dataset with an appendix stating results of testing on CelebA dataset . However, while doing our experiment, the pipeline was failing right after the GAN training. We got in touch with the authors and tried to mend the errors without success. At that point we chose to take the road less travelled which was to extract the reconstruction function from the code-base. While doing so, we had to understand the code written by authors and change it as per our requirements. After spending significant time on the fork, we successfully regenerated images that was provided by us, using the GAN . Main issues faced during this process were, figuring out proper input specifications and encoding for the images, scope of the global and local variables which were interdependent, initializing them properly and tuning the hyper parameters of the Gradient Descent based optimization algorithm. The GAN was trained till 720,000 iterations which translated to a training time of around 7 days.

While reproducing the Defense-GAN paper we reconstructed the adversarial digital images using the Generator optimization algorithm and fed the reconstructed images to the image classifier. During this operation we did the hyper parameter tuning on L and R using cross validation. Since the output of the Generator more subjective in nature (judging how well the images were reconstructed), we used a survey among our colleagues to pick the best image reproduction without revealing the parameters. As a result of this survey, we found that $L$=600 and $R$=20 gave us the best results. We fixed these hyper parameters throughout the experiments. Experimental accuracies concurred with the Defense-GAN paper. Clean images, which were both digital and physically transformed, had an average accuracy of 88%. FGSM attacks could only manage an overall average accuracy of 80% with the same falling to 65% for much coarser $\epsilon$ value such as 0.07. On finer attacks such as Carlini-Wagner and DeepFool, the accuracies were a tad higher with 82% and 84% respectively.Please refer to Table IV and V for a summary and the Appendix for detailed tables.

### D. Extension of Defense-GAN

After accomplishing the reproduction of both the Kurakin et al. [1] paper and Samangouei et al. [3] paper, we experimented with the transferability of physical adversarial images to Defense-GAN. Images transformed through the web-cam were reconstructed using the generator and fed to

TABLE IV
DIGITAL ADVERSARIAL EXAMPLES WITH AND WITHOUT GAN RECONSTRUCTION

| | *No Reconstruction* | | | | *Reconstruction* | | | |
|---|---|---|---|---|---|---|---|---|
| | **Clean** | **FGSM** | **CW** | **DeepFool** | **Clean** | **FGSM** | **CW** | **DeepFool** |
| **% Accuracy** | 100 | 0 | 0 | 0 | 87 | 65 | 82 | 87 |

the classifier. For the FGSM adversarial images that were physically transformed, the Defense-GAN achieved an overall average accuracy of 81% with the highest $\epsilon$ (0.07) images accuracy being 65%. This results proves that even after physical transformation and Defense-GAN filtering, coarse grained attacks, still manages to retain quite a bit of adversarial elements and can fool the classifier. On the other hand, CW and DeepFool images resulted in 81% and 79% accuracy after physical transformation and the Defense-GAN decreased the quality of the images. Please refer to Tables IV and V for a summary and the Appendix for detailed tables.

## IV. DISCUSSION

We were successful in reproducing the Kurakin et al. [1] paper; we showcased the effect of physical transformation on adversarial settings. Attacks such as CW and DeepFool, in comparison to FGSM attack, need very small perturbations to misclassify an input. But from our experimental results it can be seen that these finer perturbations are much easily destroyed after the physical transformation. Thus it can be inferred that finer attacks such as CW and DeepFool don't generalize very well after physical transformation. In Kurakin et al. [1] experiments, the environments were not as controlled as ours. Manually captured pictures were not aligned perfectly and there were noises introduced due to the phone not being stable. In our experiments, the webcam was kept in a fixed position in front of the LCD screen, yet we were still able to achieve a significant amount of adversarial destruction. We believe that the main reason behind this was the Moire effect. Moire effect is a visual perception effect that occurs when viewing a set of lines or dots that is superimposed on another set of lines or dots, where the sets differ in relative size, angle, or spacing. When using a camera to capture a picture of a LCD screen the pixel array of the display and pixel array of image sensor of camera must perfectly overlay to avoid Moire effect; this is impossible to achieve in practical scenarios. Thus we deduced that Moire effect acts as a major factor in the adversarial destruction of perturbed inputs.

On the Defense-GAN front, during our experiments, we could see that even after sufficient iterations and observing very low reconstruction loss, the reconstructed images were not as sharp as the input images. One reason could be the fact that the GAN itself needs much more training. We observed that samples taken after 700k iterations still had visible imperfections. Statistically speaking, test error will be equal to or higher than the training error and this could be contributing to the imperfect reconstructions on the test images. Yet it was observed that classifier gave an accuracy of 87% for original images and 65% for adversarial images created using FGSM attack ($\epsilon$= 0.07). It was also observed that on finer attacks such as CW or DeepFool the accuracies were very high. This is expected since FGSM attacks with high $\epsilon$ values produced adversarial noise which was quite prominent compared to other attack methodologies

As an extension to reproducing the results of Kurakin et al. [1]paper and Samangouei et al. [3] paper, we tested the GAN reconstructions of physically transformed adversarial examples against our classifier. We found that the the GAN continued to display low variance behavior. This means that regardless of the attack, the GAN reconstruction accuracies of physical examples seemed to remain relatively consistent. Coarse grained attacks like FGSM transcended the physical transformation better than finer attacks (CW and DeepFool). Therefore, CW and DeepFool attacks experienced much greater amounts of adversarial destruction. Therefore, using the GAN reconstruction for the FGSM physical attack resulted in a significant increase in accuracy in comparison to the FGSM examples that had not been reconstructed. On the other hand, because of the already high adversarial destruction, the use of a GAN reconstruction for CW and DeepFool examples had only a very marginal effect, even reducing the accuracies (very slightly) in some cases. Therefore, in a practical physical scenarios, it is difficult to come up with a theoretical framework to decide whether or not to use the Defense-GAN technique. From the results, we infer that generally, an optimal adversary would prefer to use FGSM (or other coarse attacks) over CW, DeepFool, or other finer attacks to craft adversarial examples for the physical world. Ultimately, we believe the choice of utilizing Defense-GAN in physical scenarios is a subjective decision; does the defender want to have the assurance of lower variance/more predictability in the accuracy of the model regardless of attack at the cost of a marginal trade-off in accuracy for certain attacks (CW/DeepFool) or does the defender want to risk high variance behavior, in terms of accuracy, and not use the Defense-GAN reconstruction. A corollary to this is that we believe that if the Defense-GAN is used in production systems, in physical environments, that the GAN should be trained for much longer. For our experiments, we trained the GAN to 700k iterations. While this was enough to demonstrate and prove our hypothesis with novel results,the reconstructions were still not completely perfect. Due to time and computational constraints, we could not train the GAN for more than 700k iterations, however, we believe that in a production environment, the GAN would be able to be trained for much longer (months instead of days). Furthermore, we believe that if the GAN was trained for that long, the 'tradeoff'

TABLE V
PHYSICAL ADVERSARIAL EXAMPLES WITH AND WITHOUT GAN RECONSTRUCTION

| | *No Reconstruction* | | | | *Reconstruction* | | | |
|---|---|---|---|---|---|---|---|---|
| | **Clean** | **FGSM** | **CW** | **DeepFool** | **Clean** | **FGSM** | **CW** | **DeepFool** |
| **% Accuracy** | 95 | 10 | 80 | 87 | 89 | 65 | 81 | 79 |

in accuracies for some reconstructions (clean/CW/DeepFool), as described earlier, would be much lower. Therefore, in this scenario, the use of the Defense-GAN technique in physical settings would be much more justifiable.

## V. CONCLUSION

In this paper, we tested the transferability of Defense-GAN to physical adversarial images. We performed this experiment using the CelebA dataset. We initially recreated the Kurakin et al. [1] paper by testing the classifier using physical adversarial examples, generated using the FGSM, CW, and DeepFool attacks. As we were focusing on the security and safety aspect of computer vision systems used in autonomous vehicles, we created physical adversarial images using a webcam to capture the adversarial images that were displayed on a LED monitor. This is in contrast to printing out images as described in the Kurakin et al. [1] paper. As expected, all attacks suffered from adversarial destruction; specifically, the CW and DeepFool attacks performed very poorly after undergoing physical transformation.

Next, we reproduced the Samangouei et al. [3] paper using the CelebA dataset. We were forced to make extensive modifications to the Defense-GAN code to allow for usage with the CelebA dataset. Ultimately, we forked a customized version of the reconstruction module in the Defense-GAN code to run our experiments. We generated reconstructions of digital adversarial images and tested these samples on the classifier. We successfully reproduced the results presented by Samangoui et al. [3]; the reconstructions of the digital adversarial images contained far less adversarial noise resulting in a higher classification accuracy.

Lastly, we present novel results by testing the effectiveness of the transferability of the Defense-GAN method to physical adversarial images. We used generated physical adversarial images, as described above using a LED monitor, and generated reconstructions using Defense-GAN. We compared the classification accuracy of physical adversarial examples with and without reconstruction. The Defense-GAN exhibited low variance behavior, producing classification accuracies generally around 80% for physical adversarial examples, regardless of adversarial attack. Therefore, we conclude that the Defense-GAN did indeed generalize well to the physical variants of adversarial examples. Furthermore, we show that the Defense-GAN works as a sensible defense for coarser adversarial attacks in the physical domain, like FGSM, since there is still high adversarial noise even after physical transformation. On the other hand, Defense-GAN may not be needed for finer/more complex attacks, like CW and DeepFool, in the physical domain, since there is already such a high rate of adversarial destruction through the physical transformation.

## VI. FUTURE WORK

Potential directions to explore with respect to our accomplished work can include in-depth analysis of Moire effect on adversarial destruction and identifying the extent to which it contributes to the destruction. As our initial motivation was focused around perturbing digital road signs, we carried on experiments by using a webcam for capturing images from a digital screen. A viable future approach to eradicate Moire effect can be printing out the images and using a camera to capture it. Thereby comparing and contrasting the effect of Moire effect on adversarial destruction.

During our experiments, we had issues with training GAN and converging properly. We were working with a GAN version termed Wasserstein GAN which used Earth Mover Distance as the metric to compare G and D distributions. However, the GAN research area is fast evolving and newer models such as EBGAN and BEGAN showcases results bettering WGAN, both in convergence and quality of the reconstructions. Experimenting with the different types of GANs looks like a promising path to take in the future.

Another feasible direction for future research could be about combating the defense GAN mechanism. Presuming the adversary can design an input image which projects onto some desired adversarial reconstructed image,that can fool the classifier which is being protected by the Generator. This would, theoretically, render the Defense-GAN useless, bringing to question its robustness.

## VII. ACKNOWLEDGMENTS

We would like to express our gratitude to Professor David Lie, for providing his patient guidance, encouragement and useful critiques to this research work. We would also like to thank Pouya Samangouei, Maya Kabkab, and Rama Chellappa, for their novel ideas and and assistance during Defense-GAN [3] paper reproduction. We would also like to thank Ian Goodfellow for opening up interesting avenues in Adversarial examples and extending the same to physical domain.

## VIII. CONTRIBUTIONS

Rohan and Bibin worked closely in setting up the development environment that was used to conduct deep learning as well as troubleshooting, customizing, and modifying the Defense-GAN code. Consequently, they also tuned and trained the GAN, the deep learning classifier, and produced the CW attack examples.

Mihir worked on generating physical adversarial examples for all attacks, as well as assisted in training the GAN and producing the reconstructed images.

Chaman worked with the FGSM attack formulation, assisted in pre-processing the images and helped in the generation of physical adversarial examples.

Karthik focused on generating the DeepFool attack examples and assisted in producing the reconstructed images using the GAN.

## REFERENCES


[1] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," *CoRR*, vol. abs/1607.02533, 2016. [Online]. Available: http://arxiv.org/abs/1607.02533

[2] A. Talebpour and H. S. Mahmassani, "Influence of connected and autonomous vehicles on traffic flow stability and throughput," *Transportation Research Part C: Emerging Technologies*, vol. 71, pp. 143–163, 2016.

[3] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-gan: Protecting classifiers against adversarial attacks using generative models," *CoRR*, vol. abs/1805.06605, 2018. [Online]. Available: http://arxiv.org/abs/1805.06605

[4] S. Chen, C. Cornelius, J. Martin, and D. H. Chau, "Robust physical adversarial attack on faster R-CNN object detector," *CoRR*, vol. abs/1804.05810, 2018. [Online]. Available: http://arxiv.org/abs/1804.05810

[5] D. Meng and H. Chen, "Magnet: a two-pronged defense against adversarial examples," *CoRR*, vol. abs/1705.09064, 2017. [Online]. Available: http://arxiv.org/abs/1705.09064

[6] P. Lu, P. Chen, K. Chen, and C. Yu, "On the limitation of magnet defense against $l_1$-based adversarial examples," *CoRR*, vol. abs/1805.00310, 2018. [Online]. Available: http://arxiv.org/abs/1805.00310

[7] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, Dec. 2015.

[8] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative Adversarial Networks," *ArXiv e-prints*, p. arXiv:1406.2661, Jun. 2014.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein GAN," *ArXiv e-prints*, p. arXiv:1701.07875, Jan. 2017.

## APPENDIX

TABLE VI
FGSM ATTACK CLASSIFIER ACCURACIES (DIGITAL AND PHYSICAL)

| | Epsilon | % Digital | % Physical | % GAN- Digital | % GAN -Physical |
|---|---|---|---|---|---|
| **Clean** | **0** | 100 | 95 | 87 | 89 |
| **Phase1** | **0.0005** | 100 | 94 | 88 | 89 |
| **Phase2** | **0.003** | 96 | 91 | 86 | 82 |
| **Phase3** | **0.005** | 86 | 87 | 86 | 85 |
| **Phase4** | **0.009** | 61 | 80 | 84 | 84 |
| **Phase5** | **0.02** | 24 | 59 | 82 | 82 |
| **Phase6** | **0.04** | 7 | 33 | 71 | 79 |
| **Phase7** | **0.07** | 0 | 10 | 65 | 65 |
| **Average** | | 53.43 | 64.86 | 80.29 | 80.86 |

TABLE VII
CW ATTACK CLASSIFIER ACCURACIES (DIGITAL AND PHYSICAL)

| | LR | %Digital | %Physical | %GAN- Digital | %GAN -Physical |
|---|---|---|---|---|---|
| **Clean** | **0** | 100 | 95 | 87 | 89 |
| **Phase1** | **0.005** | 0 | 80 | 82 | 81 |

TABLE VIII
DEEPFOOL ATTACK CLASSIFIER ACCURACIES (DIGITAL AND PHYSICAL)

| | %Digital | %Physical | %GAN- Digital | %GAN -Physical |
|---|---|---|---|---|
| **Clean** | 100 | 95 | 87 | 89 |
| **Phase1** | 0 | 87 | 84 | 79 |

Fig. 3. Digital vs Physical Adversarial Examples Classifier Accuracies Graph

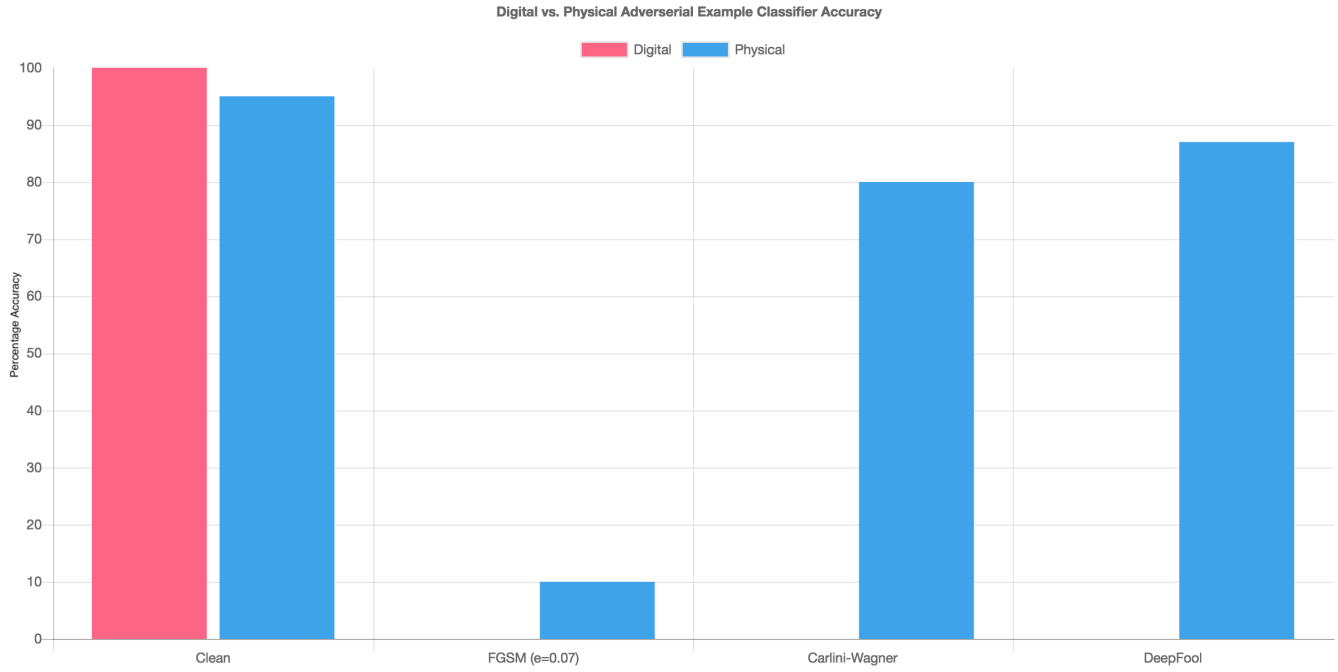


Fig. 4. Digital Adversarial Examples Classifier Accuracies with GAN vs without GAN Graph

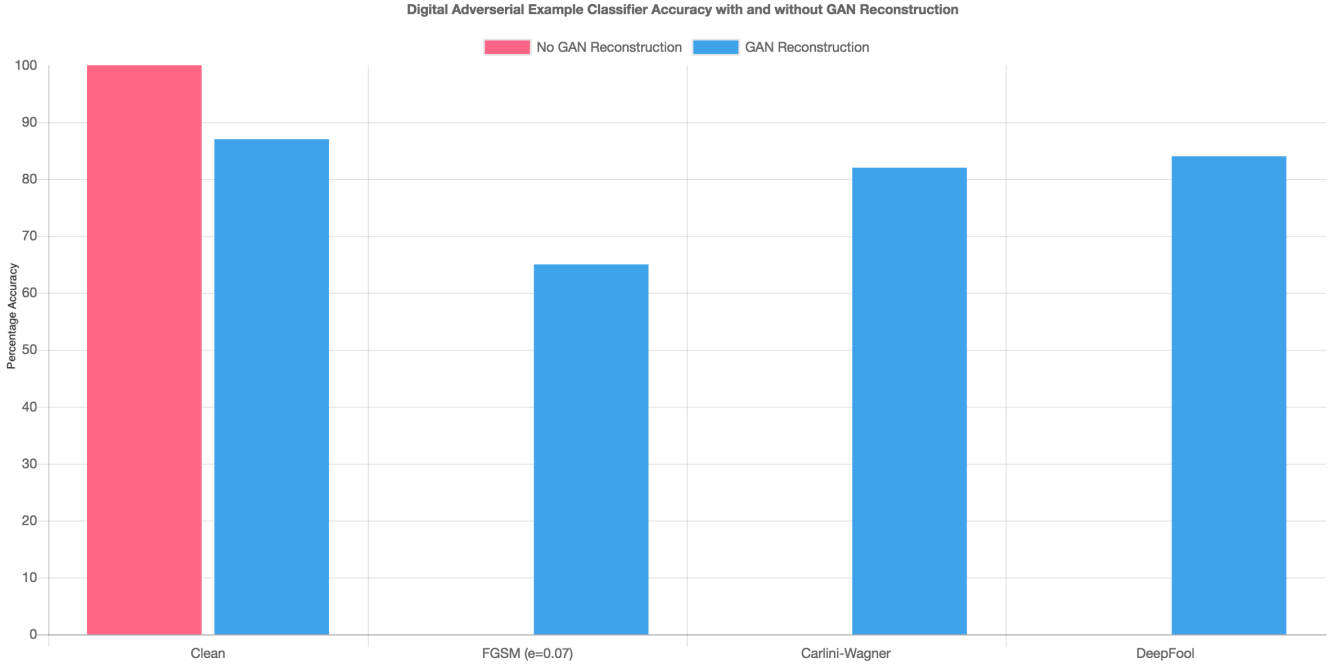



Fig. 5. Physical Adversarial Examples Classifier Accuracies with GAN vs without GAN Graph

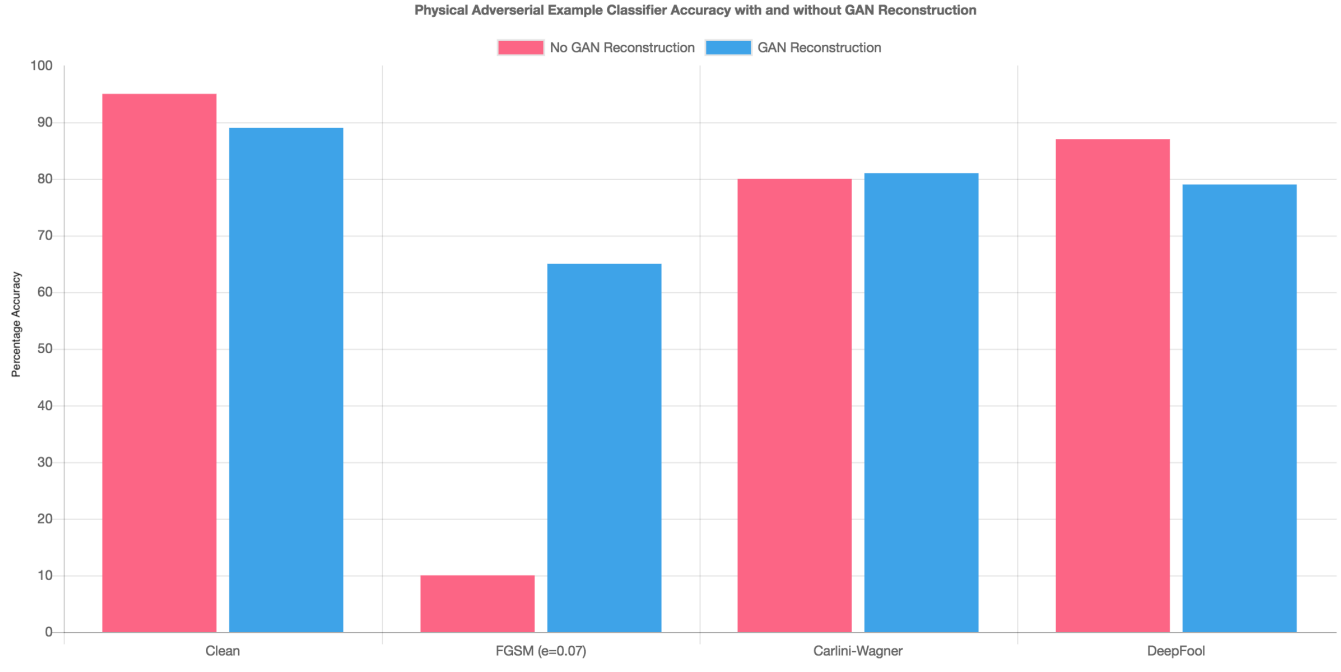


Fig. 6. FGSM Adversarial Image Samples - Digital Attack

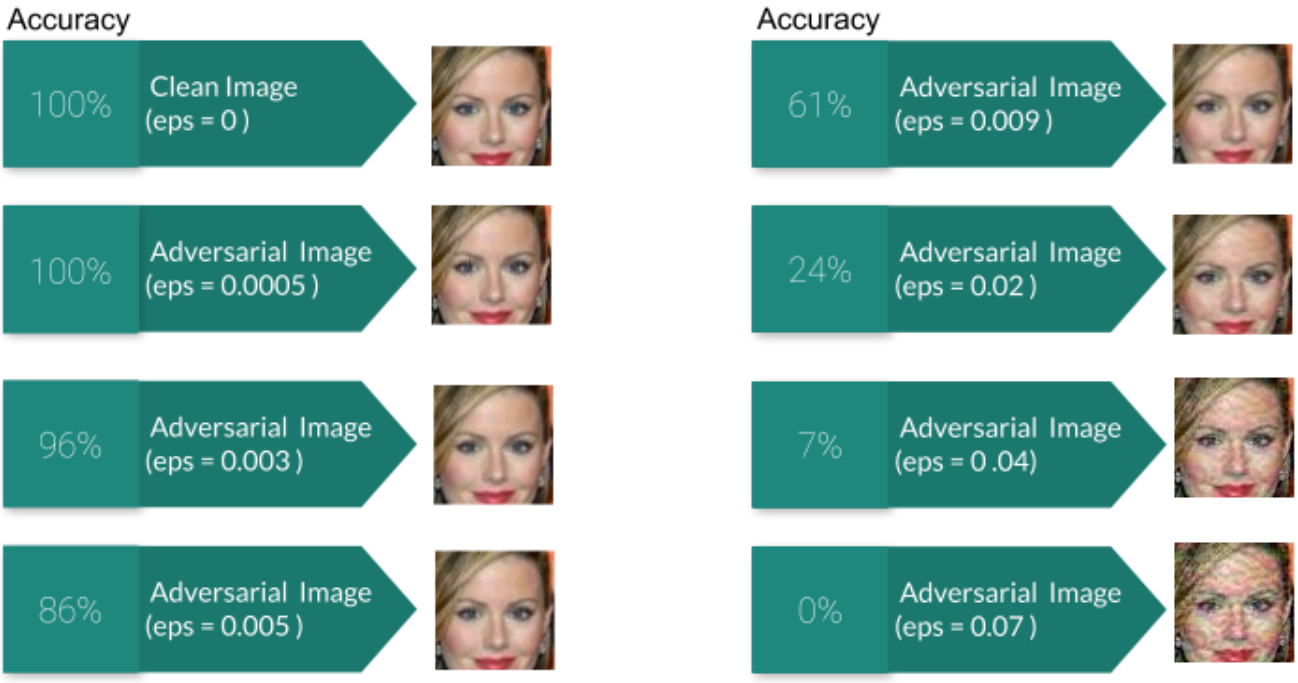



Fig. 7. CW Adversarial Image Samples - Digital Attack

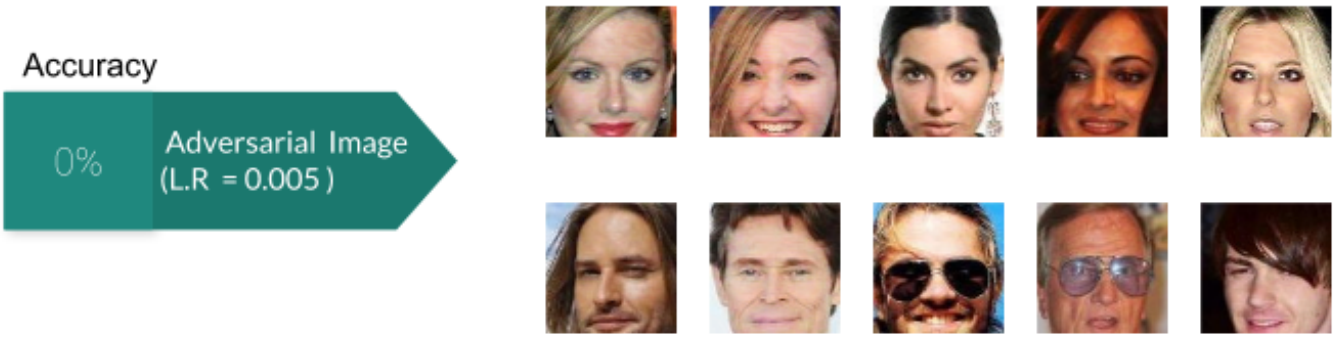



Fig. 8. DeepFool Adversarial Image Samples - Digital Attack

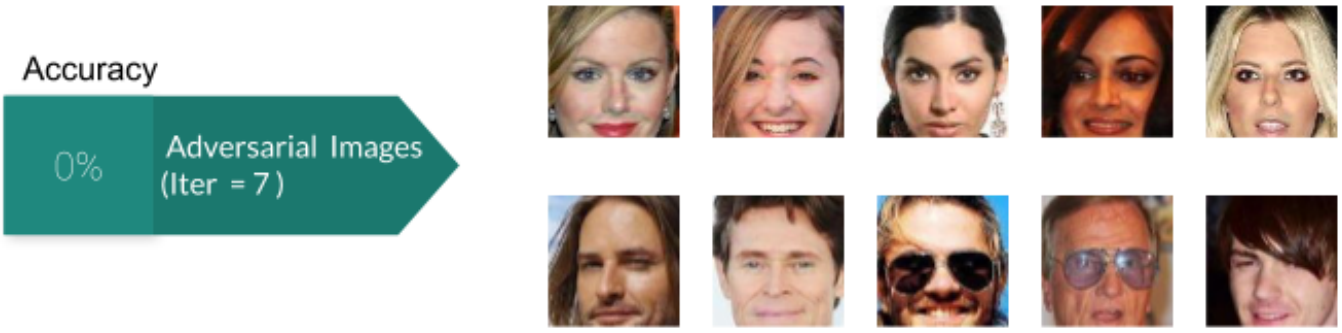


Fig. 9. Comparison of FGSM, CW & DeepFool Advesarial Images - Digital Attack

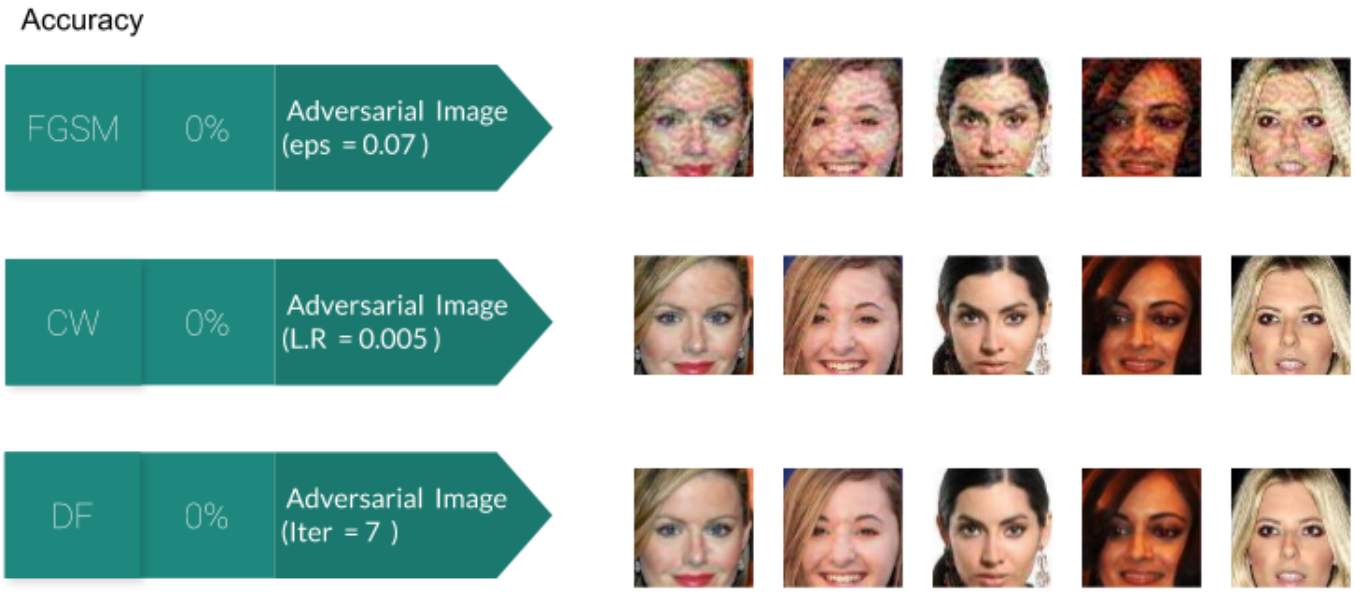



Fig. 10. FGSM Adversarial Image Samples -Physically Transformed Attack

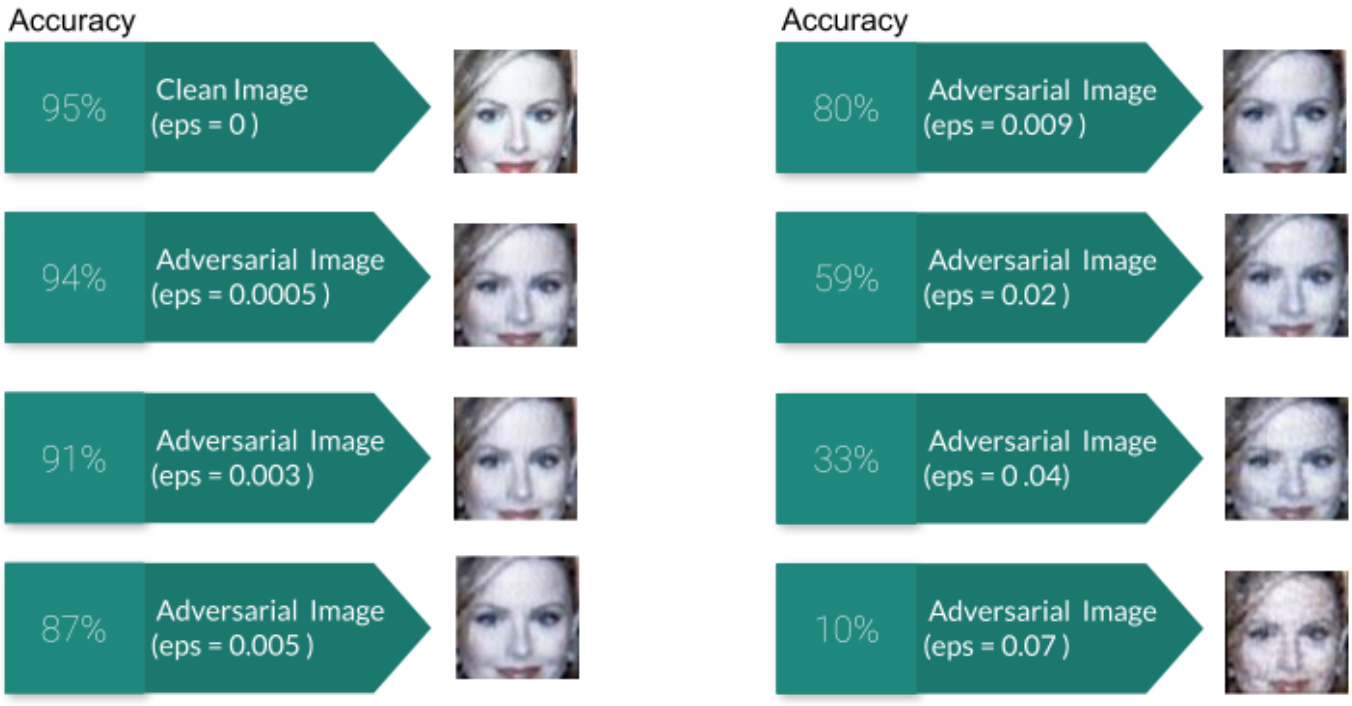


Fig. 11. CW Adversarial Image Samples -Physically Transformed Attack

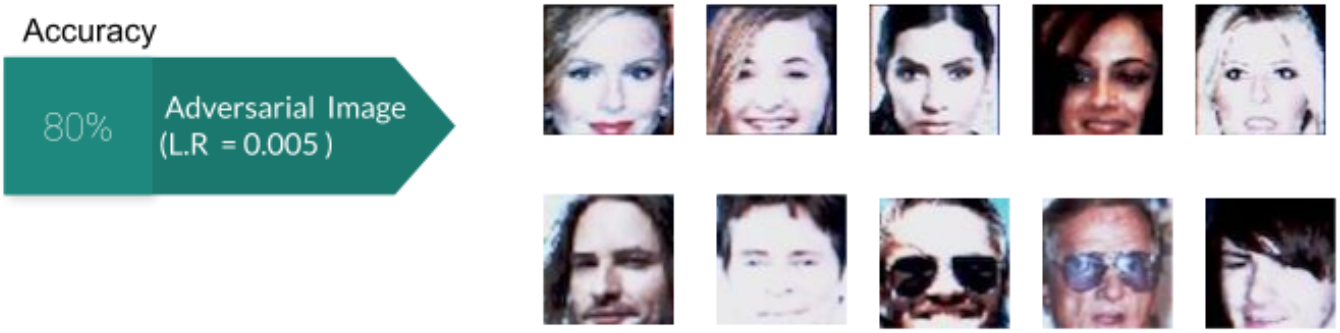



Fig. 12. DeepFool Adversarial Image Samples -Physically Transformed Attack

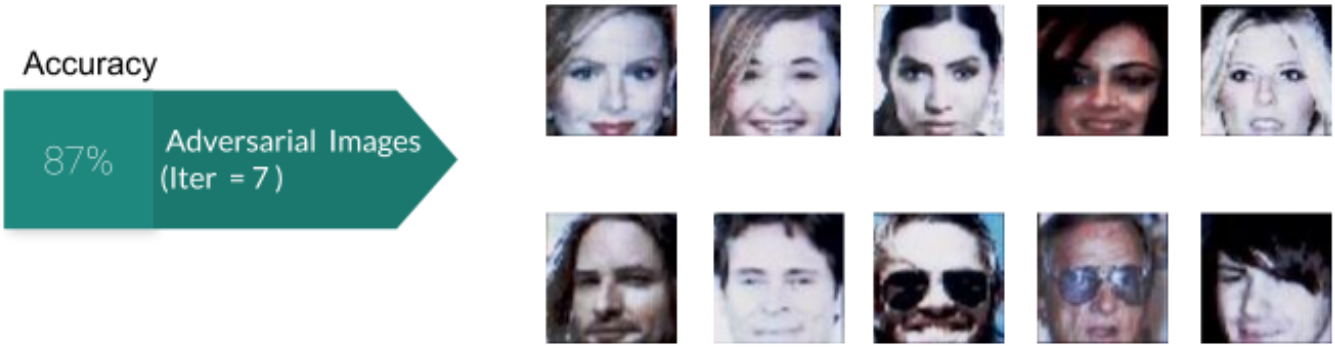



Fig. 13. FGSM Adversarial Image Samples -Physically Transformed Attack

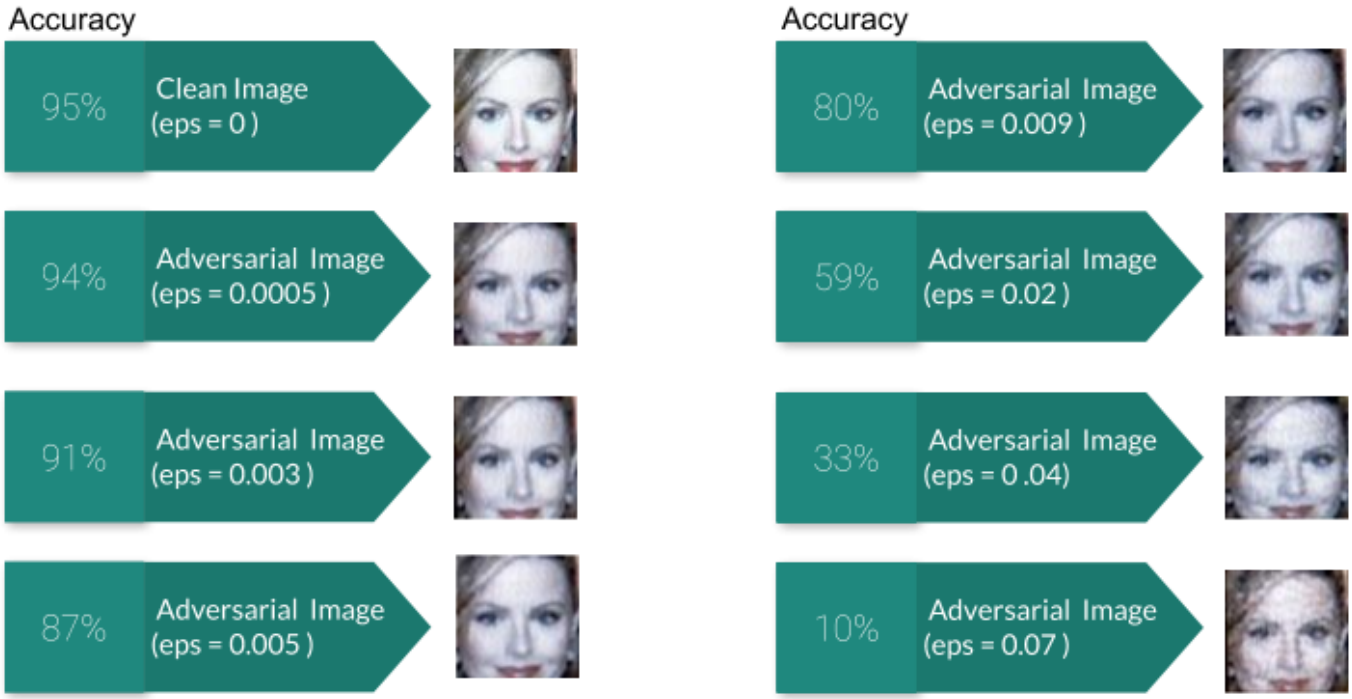


Fig. 14. Comparison of FGSM, CW DeepFool Advesarial Images -Physically Transformed Attack


Fig. 15. GAN reconstructed images with Comparison- FGSM Attack

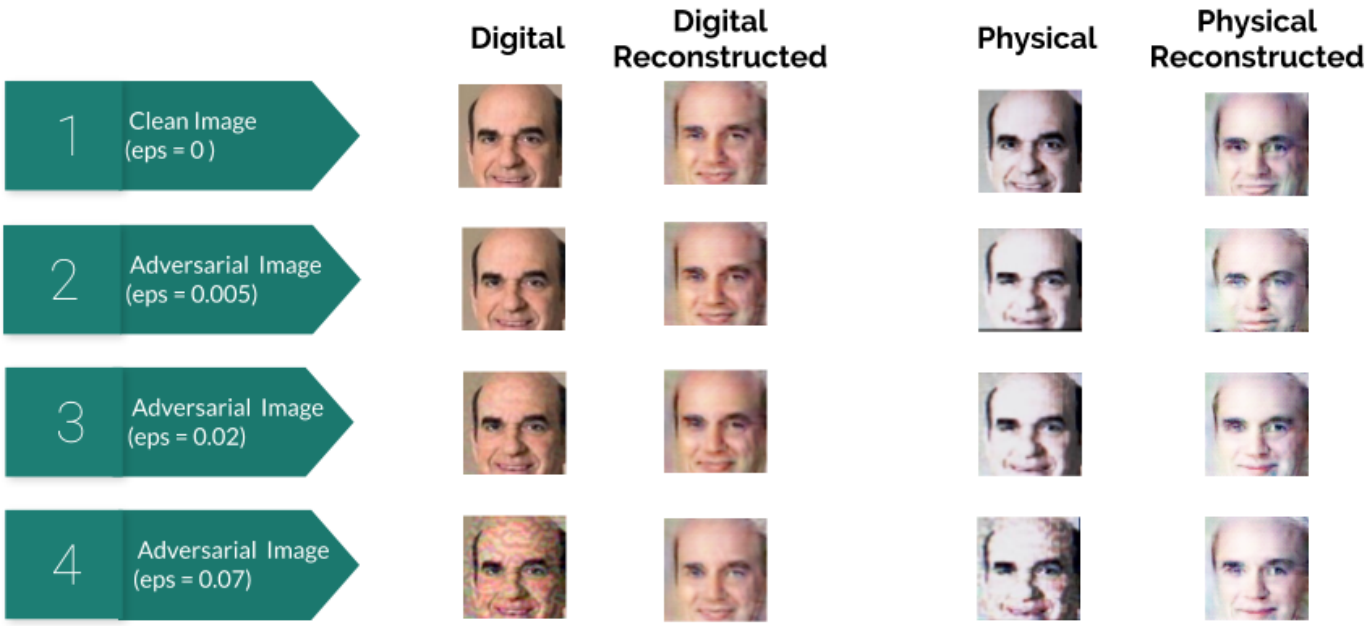


Fig. 16. GAN reconstructed images- CW Attack


Fig. 17. GAN reconstructed images- DeepFool Attack